Target selection based on fuzzy clustering: a volume prototype approach to CoIL Challenge 2000

Uzay Kaymak
Erasmus University Rotterdam
Faculty of Economics
Department of Computer Science
P.O. Box 1738, 3000 DR
Rotterdam, the Netherlands
e-mail: u.kaymak@ieee.org

Magne Setnes
Heineken Technical Services
Research and Development
Burg. Smeetsweg 1, 2382 PH
Zoeterwoude, the Netherlands
e-mail: magne@ieee.org

Abstract

A fuzzy clustering based solution to the CoIL Challenge 2000 is described. The challenge consists of correctly determining which customers have caravans in a real world customer data base, and of describing the characteristics of their profile. The solution provided uses fuzzy clustering to granulate different features and determines a score for each cluster. A version of the fuzzy c-means algorithm extended with volume prototypes and similarity based cluster merging is applied for the clustering. A score for each customer is determined from their membership to different clusters and used to select potential caravan owners. Feature selection is also performed using the scores of different clusters. This provides a transparent model that can be used for describing the profile of the potential customers.

1 Introduction

The task in CoIL challenge 2000 is a well-known data mining problem from the world of direct marketing: predict the profiles of potential customers for a product, given information about the clients and a test sample of customers possessing the particular product. Many methods have been developed for this task of *target selection*, including methods based on computational intelligence. One such method based on fuzzy clustering is described in [4]. Here we explain the application of this method for solving the CoIL Challenge 2000 task: predicting caravan owners from the client data of an insurance company. In the following, a general, but brief description of the method is given. For more details, the reader is referred to the given references.

Section 2 gives an outline of the fuzzy clustering approach to target selection. Section 3 describes the main characteristics of the data set of CoIL Challenge 2000. Section 4 describes how various experiments have been conducted to arrive at the model submitted for the challenge, followed by Section 5 that shows the results from the submitted model. Finally, Section 6 presents a discussion of the results and some learning points.

2 Fuzzy target selection

This section presents an outline for the application of fuzzy clustering in target selection. The general case is considered including all the steps that should be considered for a successful application of the method. Note that not all these steps were actually required for dealing with the specific data set of CoIL Challenge 2000. More information on the specifics of the described method can be found in [4] and [3]. Details of clustering using volume prototypes and similarity based cluster merging with an adaptive threshold are provided in [2] and [3].

2.1 Data preprocessing

In general, preprocessing of data is an important step in any data mining analysis. The most important preprocessing step for the method we have used is the removal of missing values from the data. The missing values can be removed either by removing the features that contain missing values, or by removing data patterns (client records) containing missing data. There is a trade off between preserving the number of features (variables) in the data set and preserving the number of client records. On one hand one wants to keep as many features as possible in the data set, since removing too many features may eliminate an important feature for the description. On the other hand, one wants to keep as many client

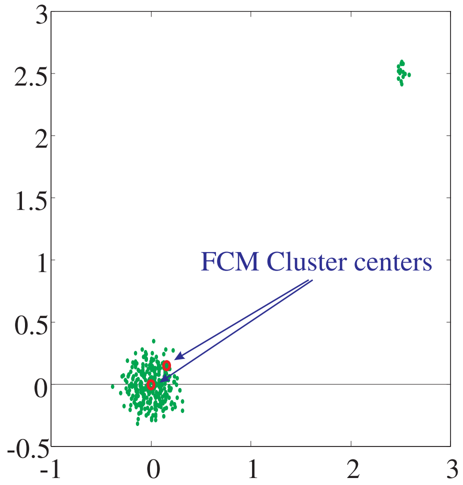

Figure 1: The location of cluster centers in fuzzy c-means is adversely influenced by a too skew distribution of data. In this example the larger cluster has 1000 points, and the smaller cluster has 15 points.

records as possible, since removing too many clients may eliminate an important client group from the data. A good way to deal with this trade off is achieved by applying a fuzzy decision making method for determining which features and which clients can be removed, while keeping as many client records and features in the data as possible. Details are described in [3].

2.2 Fuzzy clustering

Having removed the missing values from the data set, the actual modeling process can start. The basis of modeling is fuzzy clustering. The purpose of fuzzy clustering is to granulate the data such that the granules with large numbers of targeted clients (e.g. caravan owners) can be identified. Ideally, one would like to cluster in the product space of features. However, the number of features in target selection is typically large (50 to 250 features are common), and hence clustering in such a high dimensional space is computationally prohibitive. Moreover, the data then tends to be sparse (there is always a feature where two records differ), and clustering algorithms fail in dealing with such data. Therefore, it is chosen to consider one feature at a time and cluster in this one dimensional space.

Since the clustering is made in one dimension, fuzzy c-means clustering [1] is in principle sufficient. However, two problems must be dealt with. First, one needs to determine the number of clusters that are present in the data. Secondly, the data may be distributed unevenly with many points close to one another and others spread around. This influences adversely the location of the cluster centers, which may be unjustly closer to the data groups with large numbers of data points, as shown in Figure 1. The former problem is addressed by starting the clustering with a large number of clusters initially, and by merging similar clusters as the optimization progresses. The merging is based on similarity measures [2]. The latter problem can effectively be addressed by extending the fuzzy c-means algorithm with volume prototypes. The influence of the large clusters on the clustering results is then reduced.

2.3 Scoring

At this point data are clustered for all features in the data set. In other words, if there are n features, the extended fuzzy c-means clustering is performed n times. For each cluster a response density can now be computed. The response density is an indication of the predictive power of a cluster for the targeted groups. In the CoIL Challenge 2000 data, for example, this corresponds to the ratio of the number of caravan owners to the total number of people in the cluster, weighted by their cluster membership. The response density provides a measure for ordering the clusters in the order of their predictive power. Since a client can belong to more than one cluster (due to fuzzy clustering), one needs to determine a score for each client given his/her membership to different clusters and the corresponding response densities. This client score is then used for feature selection and in further modeling. Note that since the clustering is performed n times, there are also n scores for each client at this stage.

2.4 Feature selection

Because one feature is considered at a time, the basic question during modeling is the selection of the relevant features for the target groups. We follow an essentially elitist strategy here, as follows. Using the score for each client, a gain chart (sometimes also called a lift- or pareto chart) can be plotted per feature. The gain charts for different features can be compared with one another, and the feature with the most favorable gain chart is taken as the best "explanatory" feature. The clusters for the selected feature together with the corresponding response densities are recorded as part of the total target selection model.

Afterwards, the clients with highest membership to the cluster with the highest response density in the selected feature are removed from the data set. These form essentially potential targets. Then, the clustering procedure is performed again using the reduced data set and the remaining features until a pre-specified depth is reached, or until all clients are accounted for. At each stage of modeling, the selected feature, the cluster centers and the corresponding response densities are recorded as part of the target selection model.

2.5 Final model

A score is now calculated for each client for each selected feature. In the final stage, the different scores for the features are aggregated using the averaging operator. One can use a simple averaging operator, or a weighted averaging operator, where the weights can be determined as a least squares estimate. The final score calculated for each client is used to construct a gain chart and to select the potential targets. The selected features with the corresponding clusters and response densities provide a profile for the targeted customers. This profile is expressed in terms of a set of linguistic rules, as often seen in fuzzy systems.

Summarizing, the fuzzy target selection algorithm works as follows.

1. Use fuzzy decision making to remove missing values from the data set.
2. Cluster data using extended fuzzy c-means in each feature.
3. Determine cluster response densities for all clusters.
4. Score all clients for each feature using cluster membership and cluster response densities, building a gain chart for each feature.
5. Select the feature with the best gain chart. Record feature, cluster centers for that feature and corresponding response densities.
6. Reduce the data set by removing the selected feature and all client records that have their highest score in the selected feature.
7. Repeat from item 2 until the data set is empty or until a predefined number of features are selected.
8. Determine the final score for each client by averaging his/her scores for the selected features.

Figure 2 shows how the modeling proceeds in a training data set with N clients.

3 Data properties

The data of CoIL Challenge 2000 are derived from real world business data of an insurance company. The data set contains 85 features with information about demographics, possession of products and the contributions for various insurance products. The target variable is the ownership of a caravan. The goal of the challenge was to determine who the caravan owners are and to describe their profile. The data have been divided into two groups. The training data set contains 5822 clients where the labels for caravan ownership are known. The evaluation data consists of 4000 clients. The participants to the challenge had to determine, from the 4000, 800 clients whom they considered to be potential caravan owners. The data set had no missing values, and consisted of only binary and categorized features. The caravan owners form 5.9% of the total training set.

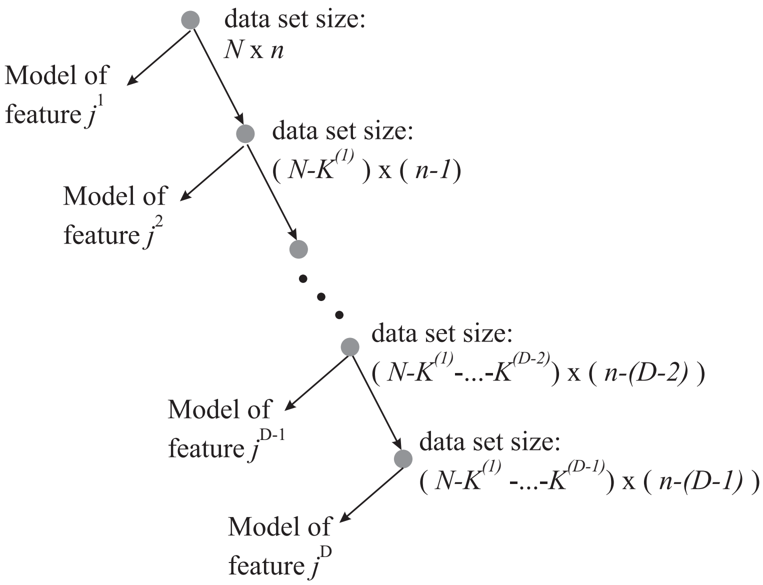

Figure 2: Fuzzy target selection for a data set with N clients and n features. At every step j of modeling one feature and $K^{(j)}$ data points are removed from the data set.

4 Experiments

Since there are not any missing values in the data, the first step of the modeling process is skipped. A three-fold cross-validation approach has been adopted for the rest of the modeling. The data set is divided randomly into three groups of equal size. Three models have been trained using two groups for training and one group for evaluation each time. Several batch runs have been performed in order to identify a setting with acceptable performance for the two important model parameters:

- number of features in the model,
- starting number of clusters for the clustering algorithm.

With a selected parameter setting, the model is trained on the total training set and then applied to the evaluation data with 4000 clients.

5 Results

Figure 3 shows the gain charts obtained with 20 features starting with 10 clusters each time and reducing the number of clusters. This model correctly predicts 108 caravan owners in the selected 800. This corresponds to a gain of 45% over the selected 20%.

For the description task the following variables turned out to be the most important. The following model has been submitted to the CoIL Challenge 2000 that the referees saw fit to reward with a special note.

1. **PPERSAUT:** Contribution car policies. Pattern 6 (fl. 1000 - 4999), response class very high.
2. **PBRAND:** Contribution fire policies. Pattern 4 (fl. 200 - 499), response class very high.
3. **APERSAUT:** Number of car policies. Pattern 2 (2), response class very high.
4. **MINKGEM:** Average income. Patterns 4 - 7 (37% - 88%), response class high.
5. **MOSHOOFD:** Customer main type. Patterns 2,9 (Driven growers and conservative families), response class high.

Furthermore, the following remarks have been noted, since any kind of derived explanatory model must also endure the test of logic. These remarks are made regarding the above patterns discovered by our model, without further knowledge of the data, the data collection mechanism and the detailed semantics of the variables recorded in the data set.

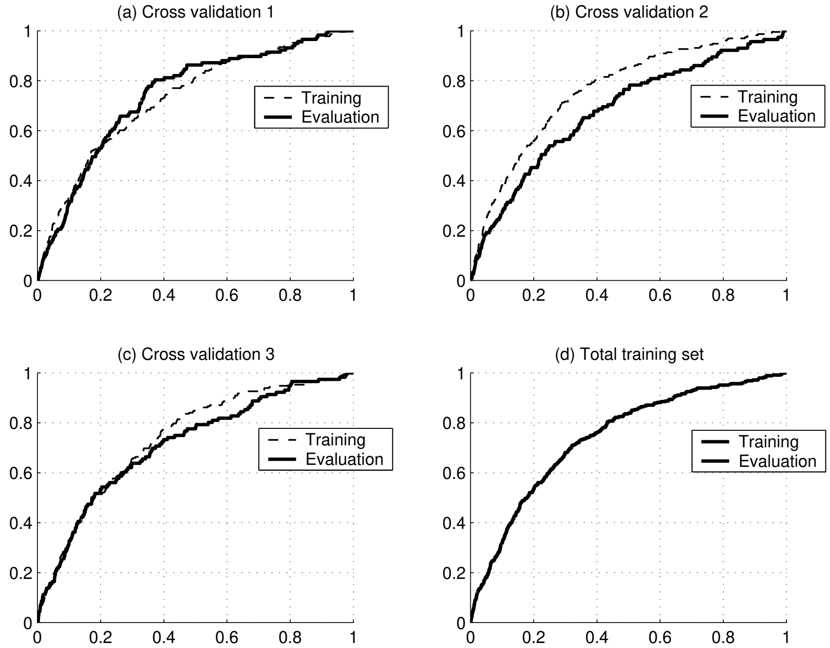

Figure 3: Gain charts obtained on the training set using a maximum of 20 features and initializing clustering with 10 clusters each time. Charts (a), (b) and (c) shows three-fold cross validation results. Chart (d) shows gain charts after training the model on the total training data set.

1. When the considered goal is the ownership of caravan insurance policy, one would expect the good predictive variables to be related to car ownership. In this sense, contribution to car policies sounds a logical choice. Note that the selected class is quite high (although rather wide). One reason will obviously be multiple car ownership as indicated by variable 3. Another reason could be that caravan owners tend to have new and medium to large cars, which would correlate well with the high car contributions.

2. In the first instance, it is not clear why the contribution to fire policies should be such an important variable. One might argue that people who have many insurance policies tend to buy more insurance policies (they constantly need assurance?) and bring them under a single company. However, this leaves the question why the contributions to other policies do not seem to be important.

   Another (and possibly more likely) possibility is that people with caravans need to store them somewhere for most of the year. After all, caravans are only used in holiday seasons and in the week-ends. In this case, the caravan may need insurance against damages like fire. If this is the case, however, PBRAND is a consequence of owning a caravan, rather than explaining who might buy caravan insurance policies, which is the ultimate goal of prediction in the marketing problem considered. Our modeling method does not use causality relations amongst variables.

3. Apparently, caravan owners tend to have two cars, possibly one small and one powerful enough to tow a caravan.

4. We have interpreted the patterns for the variable MINKGEM as meaning average to above average income. This fits well with the pattern of multiple cars and significant contribution to car insurance policies.

5. MOSHOOFD is already the result of someone's interpretation of the customer into one of several classes. It is not clear what each label means, but it would appear that ”driven growers” and ”conservative families” have a bias for caravans. May be camping with the whole family is a favorite pass-time of the conservative families, for which the caravans are used.

6 Discussion

As it can be seen from Figure 3, the selected model suffered from some over training. This is especially visible in cross validation 2. However, the final deviation (on the evaluation data set) from the predicted result is larger than expected. This suggests to us that the training data set is not representative of the evaluation data set. Nevertheless, using a simpler model

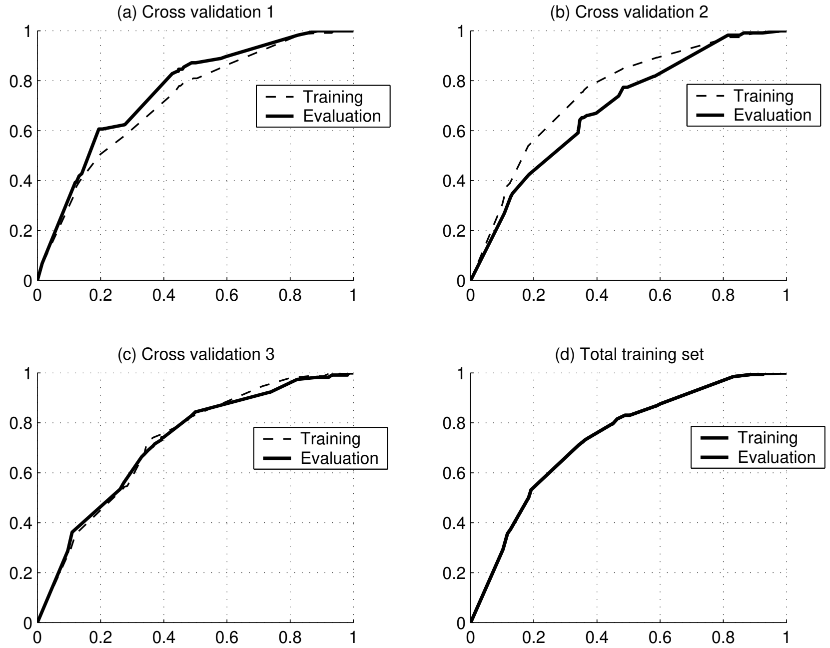

Figure 4: Gain charts obtained on the training set using three features and initializing clustering with 10 clusters each time.

would have performed better as shown in Figure 4. A model with the first three variables mentioned above (PPERSAUT, PBRAND, APERSAUT) predicts 114 of the caravan owners within the group of 800.

The CoIL data consists only of categorized and binary variables. This is unfortunate since information may be lost due to categorization as unfortunate ranges may have be selected for different categories. For example, pattern 6 for the contribution to car policies in the CoIL data ranges from fl. 1000 to 5000, which is quite wide. We are of opinion that whenever possible original(continuous) data should be used instead of pre-categorization. The clustering approach has then the potential to retrieve important information and more relevant partitioning of data.

Note that fuzzy clustering takes place in a metric space. Categorized and binary variables are essentially not suitable for use with fuzzy clustering algorithms. However, the extended fuzzy c-means clustering can deal with categorized and binary data. Because the cluster prototypes cover a volume, the algorithm does not fall over when there are only a few distinct values in the data. Instead, crisp clusters are identified and they correspond to distinct categories in the data set. Hence, for the CoIL data, which contains only categorized data, the clustering algorithm returns mostly the categories in the data. Clearly, this information is readily available, and clustering is a computationally expensive method to "re-discover" it. However, the advantage is that the fuzzy target selection method can deal with all types of data (continuous, categorized and binary), so that a single code needs to be maintained. In another real world application of the method reported in [3], we deal also with continuous data.

In the method applied only a single feature is considered at a time. More information can be utilized during clustering if clustering is performed in the product space of features. Such an extension of the proposed method will be investigated as future work.

References

[1] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function*. Plenum Press, New York, 1981.

[2] M. Setnes and U. Kaymak. Extended fuzzy c-means with volume prototypes and cluster merging. In *Proceedings of Sixth European Congress on Intelligent Techniques and Soft Computing*, volume 2, pages 1360–1364. ELITE, Sept. 1998.

[3] M. Setnes and U. Kaymak. Fuzzy modeling of client preference from large data sets: an application to target selection in direct marketing. Submitted to IEEE Transactions on Fuzzy Sets, May 2000.

[4] M. Setnes, U. Kaymak, and H. R. van Nauta Lemke. Fuzzy target selection in direct marketing. In *Proceedings of CIFEr'98*, New York, Mar. 1998.